

Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Eduardo Correa Soares

**Análises de Dados de Eletroencefalografia no
Estudo do Processamento Neurocognitivo da
Linguagem**

**Curitiba
2020**

Eduardo Correa Soares

Análises de Dados de Eletroencefalografia no Estudo do Processamento Neurocognitivo da Linguagem

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. PhD Paulo Justiniano Ribeiro Junior

Curitiba
2020

Análises de Dados de Eletroencefalografia no Estudo do Processamento Neurocognitivo da Linguagem

Eduardo Correa Soares¹

Prof. PhD Paulo Justiniano Ribeiro Junior²

Resumo

O objetivo deste trabalho é fazer uma comparação entre dois métodos de análise de dados extraídos de eletroencefalografia no estudo do processamento da linguagem e analisar suas potenciais vantagens e desvantagens. Desde os trabalhos pioneiros no estudo do processamento da linguagem, a literatura utiliza um procedimento que estabelece uma janela temporal de interesse, calcula múltiplas médias e, a partir delas, testa hipóteses com ANOVAs – procedimento chamado de extração de potenciais relacionados a eventos – em inglês, “event-related potentials”, ERPs. Uma análise alternativa é baseada na decomposição do sinal em tempo e frequência como uma série temporal, por meio da convolução entre uma *Wavelet* Complexa de Morlet e o sinal do EEG (Transformada de Fourier), seguida de um teste de hipótese paramétrico – chamada de “Time-Frequency Decomposition” – TFD. Para comparar a extração de ERPs e a TFD, neste trabalho aplicam-se ambas as análises ao mesmo conjunto de dados. Demonstra-se que, embora a TFD seja menos precisa temporalmente que a extração de ERPs, a TFD apresenta ao menos duas vantagens: (i) possibilitar a análise de mais frequências; e (ii) necessitar de menos participantes para aumentar a razão sinal-ruído.

Palavras-chave: EEG, ERP, decomposição em tempo e frequência, processamento da linguagem.

Abstract

This article aims to compare two methods to analyze the signal extracted from electroencephalography in the electrophysiology of language processing and to study their potential advantages and disadvantages. Since pioneering work in the electrophysiology of language processing, the most used procedure involves to establish a time window of interest, to multiply average the data and, based on the result, to test hypotheses with ANOVAs – a procedure called extraction of event-related potentials – ERPs. An alternative analysis is based on the decomposition of the signal in its time and frequency as a time series, through the convolution between a Morlet Complex Wavelet and the

EEG signal (Fourier Transform), followed by a parametric hypothesis testing – known as “Time-Frequency Decomposition” – TFD. In order to compare the extraction of ERPs and the TFD, in this article we applied both analyzes to the same data set. We show that, although TFD is less time-accurate than ERP extraction, TFD has at least two advantages: (i) it allows us to analyze more frequencies; and (ii) it requires fewer participants to increase the signal-to-noise ratio.

Keywords: EEG, ERP, time-frequency decomposition, language processing.

1 Introdução

Este trabalho se concentra na comparação entre dois métodos de análise de dados de eletroencefalografia (EEG) para o estudo da eletrofisiologia neurocognitiva: a extração de potenciais relacionados a eventos (ERP) e a decomposição em tempo e frequência (TFD). Em especial, foca-se na utilização desses métodos para análise de experimentos que estudam a eletrofisiologia do processamento cognitivo da linguagem. Para comparar a extração de ERPs e a TFD, aplicam-se ambas as análises ao mesmo conjunto de dados. Demonstra-se que, embora a TFD seja menos precisa temporalmente que a extração de ERPs, a TFD apresenta ao menos duas características notáveis para o estudo da eletrofisiologia do processamento da linguagem: (i) possibilitar a análise de mais frequências isolando-as; e (ii) aumentar a razão sinal-ruído com menos participantes no experimento.

Importância dos dados de EEG para o estudo do processamento neurocognitivo da linguagem é bem conhecida[1]. Até a invenção da EEG por Hans Berger na década de 1930, os métodos de estudo do cérebro eram extramente invasivos. Basicamente, o estudo de funções cognitivas de alta complexidade, como, por exemplo, o reconhecimento facial ou o processamento de linguagem, era inviável, uma vez que os métodos disponíveis, como biópsias, eram invasivos e só podiam ser realizados após a morte. A EEG possibilita o estudo em tempo real do processamento da linguagem, uma vez que os indivíduos podem realizar uma tarefa linguística enquanto seu sinal eletrofisiológico é registrado com alta precisão temporal. Assim, mesmo após quase 100 anos e o surgimento de outras técnicas minimamente invasivas (por exemplo, a ressonância magnética funcional e

¹ Aluno do programa de Especialização em Data Science & Big Data, soares_ec@yahoo.com.br.

² Professor do Departamento de Estatística - DEST/UFPR.

tractografia), a EEG continua sendo uma fonte valiosa de dados, especialmente quando se leva em conta sua precisão temporal em comparação a de outras técnicas [2, 3, 1].

Dada a possibilidade de conduzir a coleta com seres humanos vivos, desde o artigo de Marta Kutas e Stephen Hillyard[4], experimentos utilizando EEG são utilizados por pesquisadores para estudar diferentes fenômenos relacionados ao processamento neurofisiológico da linguagem.³ Nesse artigo fundacional em que descobriram o N400 (detalhes a seguir), Kutas e Hillyard utilizam a EEG para a extração de ERPs a partir de estímulos linguísticos apresentados no paradigma “oddball”, *i. e.*, uma sequência de estímulos em que uma das condições é especialmente diferente e aparece ocasionalmente (em proporção diferenciada, como, por exemplo, 30/70). Para esse tipo de estudo, costuma-se apresentar estímulos linguísticos escritos ou auditivos e analisar o sinal eletrofisiológico a partir de uma palavra específica que apresenta as condições, como, por exemplo, uma palavra “esperada” vs. uma palavra “anômala”, como nos exemplos da Tabela 1.

| P1 | P2 | P3 | P4 | P5 | P6 | P7 | condição |
|----|-------|--------|-------|--------|------|---------|------------|
| A | pizza | estava | muito | quente | para | chorar. | (anômala) |
| A | pizza | estava | muito | quente | para | comer. | (esperada) |

Tabela 1: Amostra do paradigma *oddball* original, adaptado de artigo fundacional de Kutas e Hillyard [4] – P representa “palavra”, apresentada em intervalos regulares de tempo, na ordem dada pelo número.

Normalmente, centenas de frases semelhantes (geralmente chamados de “itens” experimentais) são apresentadas aos participantes dos experimentos, em meio a frases distratoras, enquanto seu sinal analógico é capturado por eletrodos, enviado a um amplificador e registrado digitalmente em um computador especialmente reservado para esse registro (normalmente, rodando somente o *software* de coleta para evitar qualquer efeito temporal). Um desenho experimental comum é a utilização da distribuição de quadrado latino: cada participante lê/escuta uma condição de cada item, de modo que todos os itens sejam vistos mais de uma vez em cada condição por diferentes participantes (na proporção n^o de participantes dividido pelo n^o de condições). Os dados são tratados por diferentes procedimentos analíticos de modo a aumentar a razão sinal/ruído e, por fim, um teste estatístico de hipótese é utilizado para verificar se as condições A e B (esperada vs. anômala, por exemplo) são significativamente diferentes. Kutas e Hillyard identificaram uma deflexão negativa com uma amplitude após as palavras

“anômalas” significativamente diferente (maior) do que das palavras “esperadas”[4].

A seguir, analisam-se alguns procedimentos para o tratamento dos dados de EEG na eletrofisiologia da linguagem. Especificamente, há dois grandes grupos de análises: o procedimento clássico, aqui denominado Extração de ERPs [4, 2], e um procedimento alternativo mais utilizado em outras áreas da eletrofisiologia cognitiva, aqui chamado de TFD [3]. É importante salientar que a análise de EEG é multidimensional e discutir todos os aspectos estatísticos da modelagem desse tipo de dados estaria além do escopo deste artigo. O escopo aqui delimitado restringe-se a métodos de modelagem que são utilizados para teste de hipótese, deixando modelagens que são voltadas para análises exploratórias e predição para pesquisa futura [3].

2 Revisão de Literatura

Neste trabalho, comparam-se duas metodologias utilizadas para análise de dados de EEG em eletrofisiologia neurocognitiva. Nesta seção, apresentam-se brevemente a análise mais comum na área de neurociência da linguagem (a extração de ERPs) e uma análise alternativa, que, embora não seja ainda amplamente adotada na área de processamento da linguagem, já está bem consolidada em outras áreas da eletrofisiologia cognitiva (Decomposição em Tempo e Frequência).

2.1 Extração de ERPs

O procedimento mais utilizado para o estudo eletrofisiológico do processamento da linguagem é conhecido como extração de ERPs [4, 5, 6, 2]. A ideia essencial é que estímulos linguísticos elicitam potenciais elétricos em dipolos cerebrais (conjuntos de células neurológicas), que são então capturados pela EEG como deflexões no sinal. Essas deflexões são conhecidas como “potenciais relacionados a eventos”, pois são potenciais elétricos gerados pelo fluxo de elétrons nos dipolos associados a eventos específicos (os estímulos de um experimento, por exemplo). Nessa técnica, janelas temporais para cada evento (conhecidas como “épocas”) são estabelecidas. Essas épocas são posteriormente analisadas em função das condições, como o *onset* temporal do aparecimento das palavras “chorar” e “comer” no exemplo da Tabela 1. O procedimento pode ser sumarizado nas seguintes etapas: (i) pré-processamento dos dados e remoção de artefatos (como o piscar dos olhos); (ii) recorte das janelas temporais associadas aos eventos (dado que o sinal de EEG é em geral contínuo); (iii) opcionalmente, subtrae-se um valor de linha de base (a média da amplitude de uma janela de 200ms antes do *onset* do estímulo) de cada ponto observado; (iv) cálculo de médias entre “trials” em cada condição de um mesmo participante do experimento; e (iii) análise estatística inferencial utilizando geralmente ANOVAs de medidas repetidas, com variação entre participantes. A seguir, detalham-se os procedimentos (iv)

³ Os experimentos para investigar a linguagem são, na verdade, um pouco anteriores à publicação do artigo de 1980 [4, 2]. De acordo com Kutas, ela e Hillyard procuravam uma maneira de fazer um experimento que, em vez de apresentar palavras como listas (de nomes estereotipicamente masculinos ou femininos), apresentasse frases, esperando que o efeito fosse similar ao de listas. O que eles descobriram foi que, em frases, havia um efeito único, diferente da apresentação de listas de palavras e de outros estímulos não-linguísticos, inaugurando assim uma área de investigação: a eletrofisiologia do processamento linguístico.

e (v), bem como a visualização dos dados, que serão o foco da comparação feita neste artigo.

2.1.1 Cálculo de médias

O sinal de EEG é naturalmente ruidoso. O ruído do sinal advém de múltiplas origens, desde a atividade cerebral relacionada a diversos fatores internos ao participante do experimento, como por exemplo o batimento cardíaco e a respiração, até os fatores externos, como o ambiente em que o experimento é conduzido. Os equipamentos modernos já atenuam uma boa parte desse ruído, e o pesquisador conduzindo o experimento tenta também controlar as variáveis de modo a reduzi-lo. Para a preparação final dos dados e visualização dos ERPs, o pressuposto básico é de que o cálculo de médias cancela a variação aleatória não-relacionada aos estímulos (ruído), preservando assim somente o sinal relacionado ao evento de interesse.

O procedimento de cálculo de médias consiste em estabelecer uma janela temporal fechada no evento de interesse e em calcular médias entre todas as observações, sumarizando diversos *trials* por condição por cada participante do experimento. Suponha-se, então, que a janela temporal de interesse seja de 1000 milissegundos em um sinal de EEG gravado a uma taxa de amostragem de 500 Hertz (Hz); suponha-se, também, que cada participante foi exposto a 80 *trials* de cada tipo de estímulo. Serão realizadas 500 médias entre os 80 *trials*, *i. e.* uma média para cada ponto temporal digitalizado pelo equipamento. O pressuposto básico para que esse procedimento preserve o sinal é que o ERP é ao mesmo tempo fechado no tempo e fechado na fase, *i. e.* que haja absoluta sobreposição da latência temporal dos potenciais relacionados aos estímulos em cada *trial*. Assume-se, portanto, que os ERPs em grande medida se sobrepõem entre *trials* e o ruído é distribuído aleatoriamente com relação ao evento de interesse, às vezes positiva, às vezes negativamente [2]. Assim, a deflexão do sinal associada ao estímulo na condição A que está presente no *trial* 1 se repete no *trial* 2, 3, *etc* e será preservada na média, enquanto o ruído do *trial* 1, que não está presente no *trial* 2, 3, *etc* será cancelado na média. Note-se também que não-raro pesquisadores reportam o cálculo da diferença entre a média da condição A e B, pois essa diferença cancelaria ainda ruídos de relacionados a outras ondas, possivelmente gerados em outros dipolos.

2.1.2 Visualização no Domínio Temporal

A maioria dos pesquisadores que trabalham com a Extração de ERPs reportam gráficos representando a amplitude dos potenciais ao longo do tempo, a partir de médias globais entre todos os participantes nas condições experimentais em análise. O procedimento é idêntico ao reportado acima, porém, em vez de calcular a média entre *trials*, calcula-se a média entre participantes para cada milissegundo. Suponha-se, novamente, que a janela temporal de interesse seja de 1000 milissegundos em um sinal de EEG gravado a uma taxa de amostragem de

500 Hz; suponha-se, também, que o experimento contou com 20 participantes. Serão realizadas 500 médias para cada ponto digitalizado pelo equipamento, entre os 20 participantes para cada condição experimental. Esse procedimento geralmente leva a um sinal bem suavizado, com semi-ondas de cerca de 100 a 300 milissegundos de latência.

2.1.3 Teste de Hipótese

Tipicamente, pesquisadores utilizam ANOVAs de medidas repetidas para teste de hipótese em uma análise a nível de grupo. Supondo que todos os participantes do experimento pertencem ao mesmo grupo amostral, em uma dada latência de interesse, pesquisadores testam a probabilidade da rejeição da H_0 quanto à diferença de amplitude entre condição A e B (teste F). A maioria dos estudos com extração de ERPs se baseia em valores-p e no limite de significância de 0.05, com raros casos em que *Cohen-d scores* são usados. Normalmente, esse teste de hipótese inclui também eletrodos e, em alguns casos, tempo: em alguns estudos, utilizam-se *clusters* de eletrodos e, de acordo com uma recomendação geral da área, deve-se realizar uma média da amplitude na latência de interesse para o teste de hipótese [2, 7], o que é conhecido como "voltagem janelada".

2.2 Abordagem baseada em TFD

Uma abordagem alternativa que vem gradualmente ganhando espaço nas publicações sobre eletrofisiologia neurocognitiva é a decomposição em tempo e frequência (TFD). Embora já existam alguns trabalhos no campo do processamento da linguagem [8, 9, 10], essa abordagem ainda não se tornou padrão e não há notícia de que tenha já sido aplicada em qualquer estudo feito no Brasil. A seguir, apresenta-se brevemente a TFD quanto ao procedimento de preparação dos dados e sua visualização e quanto ao teste de hipótese, que serão posteriormente comparados à Extração de ERPs.

2.2.1 Decomposição em Tempo e Frequência com Wavelet Complexa de Morlet

Há diferentes implementações da TFD, que divergem em detalhes que não são cruciais para os objetivos deste trabalho.⁴ Este trabalho se concentra em uma das propostas de implementação da TFD cuja característica principal é a convolução entre o sinal de EEG e a *Wavelet* Complexa de Morlet, também conhecida como função

⁴ Entre as diferentes propostas de implementação de TFD, destacam-se Filtro de Banda – conhecido como Transformada de Hilbert, Transformada Rápida de Fourier restrita a uma janela temporal, Modelagem Autorregressiva, entre algumas outras. As implementações mais utilizadas divergem basicamente em três pontos da que é detalhada neste artigo: (i) o modo como lidam com a pressuposição de estacionariedade da Transformada de Fourier, (ii) como lidam com a incongruência de fase do sinal, e (iii) se utilizam um “template” de onda para a extração do sinal analítico ou não. Há vantagens e desvantagens em cada uma dessas implementações, que não afetam os resultados da análise feita neste trabalho e, portanto, não serão detalhadas aqui.

Gabor. *Wavelets* são “templates” de ondas senoidais de diferentes frequências, gerados conforme a equação 1:

$$\text{Senoide} = A \sin(2\pi f t + \theta). \quad (1)$$

Na equação, A é a amplitude da onda senoidal, π é pi (3,141...), f é a frequência da onda senoidal e t é tempo. Por fim, θ é a fase da onda, que é um parâmetro importante para a correta utilização do transformada de Fourier, como se salienta a seguir.

A particularidade da *Wavelet* de Morlet é que ela é gerada a partir da multiplicação ponto por ponto de uma onda senoidal de uma determinada frequência e de uma “janela” Gaussiana. A janela Gaussiana é gerada conforme as equações 2 e 3.

$$\text{Gaussiana} = A e^{(t-m)^2/2s^2}, \quad (2)$$

$$s = \frac{n}{2\pi f}. \quad (3)$$

Na equação 2, A é a amplitude (ou a altura da gaussiana), t é tempo, m é o offset do eixo x (que é irrelevante para a análise de EEG e, portanto, pode ser fixado em 0), s é o desvio padrão (ou a largura da gaussiana). s , por sua vez, é definido na equação 3, em que f é a frequência (em hertz) e n é o número de ciclos na *Wavelet*. O resultado da *Wavelet* de Morlet (ou seja, o formato do “template”) se assemelha à onda na Figura 1.

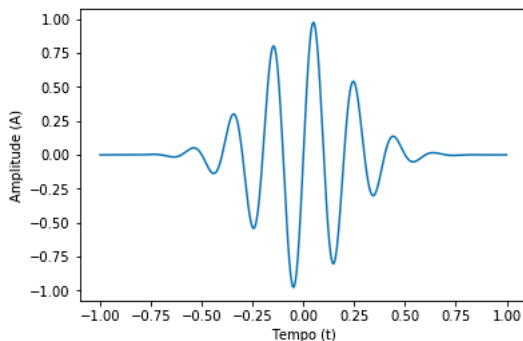


Figura 1: Wavelet de Morlet gerada a partir da multiplicação ponto por ponto de uma onda senoidal e uma janela gaussiana.

A ideia básica dessa TFD consiste em convolucionar os dados de EEG com *Wavelets* Complexas de Morlet. Essa operação consiste em aplicar a Transformada Rápida de Fourier (soma dos produtos internos dos vetores – dados e *kernel*) – entre os dados (o sinal de EEG) e um *kernel* (a *Wavelet* Complexa de Morlet). Por meio da Transformada Rápida de Fourier, extraem-se os coeficientes de Fourier para cada frequência: (i) a magnitude (M) e (ii) o ângulo (θ). É importante salientar que, diferentemente da *Wavelet* de Morlet, a *Wavelet* Complexa de Morlet é tridimensional, pois aos eixos x e y se adiciona um eixo de números imaginários para capturar a fase dos dados (a equação é a mesma com exceção de que a onda senoidal é criada com um número complexo – com uma parte real

e uma imaginária). Assim, utilizando-se da fórmula de Euler e de identidades trigonométricas, conclui-se que se pode representar o resultado da Transformada Rápida de Fourier com a *Wavelet* Complexa de Morlet por um vetor de Coeficiente de Fourier, conforme a equação 4:

$$\text{Coeficiente de Fourier} = M_t e^{i\theta_t}. \quad (4)$$

Com os coeficientes de Fourier, é possível reconstruir a série temporal por meio da Transformada Inversa de Fourier, resumizada na equação 5:

$$x_k = \sum_{k=1}^n X_k e^{i2\pi f(k-1)n^{-1}}. \quad (5)$$

Na equação, x_k representa a série temporal, X_k representa os coeficientes de Fourier. O exponencial é uma forma de reconstruir a onda senoidal. O trabalho de decomposição, portanto, se resume a iterar a Transformada Rápida de Fourier sobre múltiplos *kernels*, ou seja, sobre diferentes frequências da *Wavelet* Complexa de Morlet.

2.2.2 Visualização no domínio do tempo e da frequência

Uma das características da decomposição espectral de uma onda em seu poder (quadrado da magnitude) por frequências é que o poder de frequências mais altas é sempre menor que o de frequências mais baixas, conforme a Figura 2.

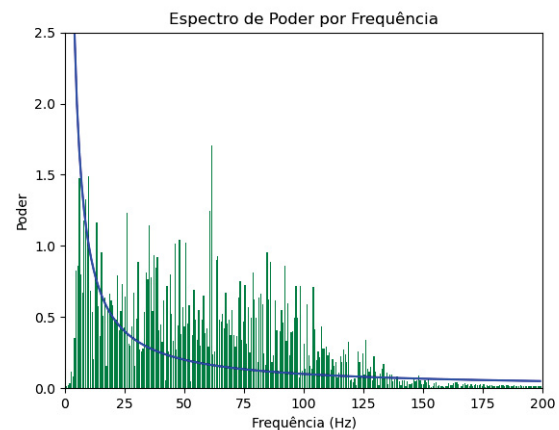


Figura 2: Distribuição Espectral do Poder a partir da decomposição em tempo e frequência. A linha representa a distribuição teórica $1/f$, enquanto as barras representam a decomposição real de 500ms de sinal de EEG.

Na Figura 2, a linha azul representa uma distribuição teórica do poder por frequências, enquanto as barras verdes correspondem à decomposição de dados reais de um *trial* de EEG com a transformada de Fourier de uma janela temporal de 500ms. Observe-se que essa distribuição teórica é própria do sinal eletrofisiológico [3].

Para a visualização da amplitude de altas e baixas frequências no domínio do tempo e da frequência em

uma única figura, há três recomendações: (i) restringir o eixo y a uma determinada banda de frequências [8], (ii) usar uma escala logarítmica no eixo y; ou (iii) fazer uma normalização da atividade após o estímulo (AAE) por frequências em uma escala logarítmica, como de decibéis, ou em z-scores ou ainda em porcentagem de mudança [3]. Essa normalização em decibéis toma a média do poder de uma linha de base (LDB) (de 100 a 200ms antes do *onset* do estímulo) como denominador e coloca em uma escala logarítmica multiplicada por 10 (por isso, “deci-”), cf. equação 6.

$$dB_{tf} = 10 \log_{10} \left(\frac{AAE_{tf}}{LDB_f} \right) \quad (6)$$

Na Figura 3, observa-se que a utilização dessas recomendações possibilita visualização de diferentes frequências. Muitas vezes, os pesquisadores optam por combinar duas ou até as três recomendações, como no terceiro painel da Figura 3.

2.2.3 Testes de Hipótese

Embora diferentes métodos na literatura sobre TFD têm sido propostas e utilizadas em eletrofisiologia cognitiva [3, capítulo VI], no estudo do processamento da linguagem as análises inferenciais utilizadas são em geral mais conservadoras. Testes-*t* bilaterais de condições pareadas e ANOVAs de medidas repetidas ainda são o procedimento mais utilizado [8]. Nesse ponto, o procedimento utilizado na TFD não difere muito da extração de ERPs: supondo que todos os participantes do experimento pertencem ao mesmo grupo amostral, em uma dada latência de interesse, pesquisadores testam a probabilidade da rejeição da H_0 quanto à diferença de poder entre condição A e B (teste F). Esse teste de hipótese pode também incluir eletrodos e, em alguns casos, tempo (médias entre eletrodos ou pontos no tempo também são utilizadas). Em casos em que há múltiplas comparações, algum tipo de correção para garantir o nível conjunto de significância é normalmente utilizada como, por exemplo, a de Bonferroni.

3 Método e Materiais

O método utilizado neste artigo consistiu em aplicar ambas as análises ao mesmo conjunto de dados. É importante salientar que o pré-processamento dos dados foi realizado somente uma vez, pois, em ambos os métodos, os procedimentos são muito similares. Foi, então, aplicado um filtro de passagem de banda alta (de 0,1 Hz), para correção do *drift*, Análise de Componentes Independentes (ICA) e inspeção visual para remoção de artefatos, como piscadas de olhos, que são facilmente identificáveis visualmente. Realizaram-se esses procedimentos com o pacote MNE para Python [11]. Por fim, todos os dados foram re-referenciados com respeito a média global de todos os eletrodos, para evitar distorções globais específicas ao participante e para centralizar

o sinal [2, 3, 7]. No restante das análises apresentadas nas seções 3 e 4, seguiram-se os procedimentos descritos na seção 2.

3.1 Conjunto de Dados

Este trabalho utiliza uma base de dados disponibilizada eletronicamente para o estudo do N400 [12]. Como salientado na seção 1, o N400 foi o primeiro e é o mais estudado (portanto, mais consolidado) efeito eletrofisiológico associado ao processamento da linguagem em EEG. Esses dados foram coletados usando uma tarefa de julgamentos de pares de palavras [13, 12]. Em cada *trial*, uma primeira palavra em vermelho era apresentada antes de uma palavra-alvo verde. Os participantes tinham de responder se a palavra-alvo estava semanticamente relacionada ou não com a palavra antecedente em vermelho apertando botões de um controle.

A base compreende dados de 40 participantes do experimento, todos neurologicamente saudáveis. Foi gravado o sinal de EEG contínuo de 30 eletrodos, com os 120 eventos de interesse marcados em pontos temporais específicos. Dos 120 eventos, 60 eram palavras em vermelho (chamadas na literatura de *prime*) e 60 eram palavras-alvo, que compuseram, portanto, 60 *trials*. Dos 60 *trials*, 30 eram compostos por palavras relacionadas e 30 por palavras não-relacionadas. Foram criadas 2 listas para contra-balanceamento dos itens em quadrado latino.

3.2 Recursos Computacionais

Cabe ressaltar que maioria das análises de dados de EEG são realizadas em *Matlab* [14] ou por softwares proprietários, vendidos ou não juntamente com o equipamento. As análises dos dados aqui reportadas foram, no entanto, realizadas em *Python* pela disponibilidade de recursos e por ser um *software* de código aberto [15]. É importante, por exemplo, mencionar que o Transformada Rápida de Fourier do pacote *scipy* é otimizada em *Python* por meio de cálculos matriciais (Álgebra Linear) e implementação em linguagem de nível baixo (C) [16]; por isso, ela, assim como boa parte dos demais recursos utilizados aqui, têm alto desempenho computacional e baixo custo. Por essa razão, *Python* foi escolhido para a análise feita aqui.

3.3 Métodos

A metodologia deste trabalho consiste em aplicar a extração de ERPs e a análise baseada em TFD à base de dados descrita na seção 3.1. Após a aplicação dos dois procedimentos, comparam-se as análises e se analisam suas vantagens e desvantagens articulando-as com a literatura recente sobre a eletrofisiologia do processamento da linguagem.

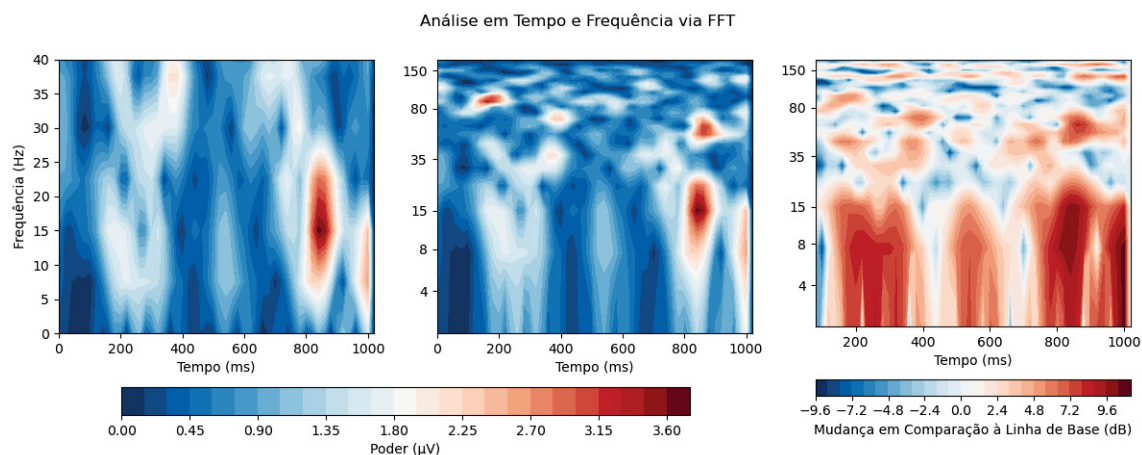


Figura 3: Comparação entre diferentes escalas de visualização do domínio de tempo e frequência. No primeiro quadro, foi reduzida a janela de frequências às visualizáveis na escala do poder; no segundo quadro, foi utilizada uma escala logarítmica para o eixo y; no terceiro quadro, os valores do poder foram transformados em decibéis.

4 Resultados

Foi selecionado um único eletrodo (Cz) para a comparação entre os métodos. Nesse eletrodo, geralmente os efeitos do N400 são bem pronunciados, o que perfaz uma hipótese bem-conhecida para teste. A hipótese a ser testada é, então, de que a condição relacionada e a condição não-relacionada diferem na janela referente ao N400, com a condição não-relacionada levando a uma deflexão negativa saliente em comparação à condição relacionada.

A princípio, foi mantida a mesma análise estatística inferencial em ambas as análises, de modo que fosse idêntica a comparação no pré-processamento, como salientado na seção 3, e no procedimento final de análise. As duas análises diferem, portanto, em (i) processamento para o aumento da razão sinal/ruído e (ii) visualização dos dados. Como consequência, ainda que seja aplicada ANOVA de medidas repetidas em ambas as análises, os valores que entraram no cálculo são diferentes.

4.1 Visualização e Análises Possíveis

Seguindo a metodologia da extração de ERPs descritas na seção 2.1, foram calculadas médias ponto-por-ponto entre *trials* por condição para cada participante. Essas médias nos levaram a 40 séries temporais de 1.5 segundo a 1024Hz (1536 pontos de amplitude). Para a visualização, foram calculadas médias globais entre todos os participantes, conforme pode ser visto na Figura 4. Adicionalmente, para lidar com o possível problema da sobreposição de pontências advindos de diferentes aspectos cognitivos, foi calculada a diferença entre as médias da condição não-relacionada e relacionada ponto por ponto, que foi plotada em verde na Figura 4. Por fim, para o teste de hipótese, seguindo a literatura [2, 7], foram feitas médias das janelas temporais de interesse por participante (de 300 a 500ms). Estas são resumidas

na Tabela 3 e nas linhas tracejadas com as respectivas cores das condições relacionada e não-relacionada na Figura 4.

| Condição | Média | DP | EP |
|-----------------|---------|--------|--------|
| relacionada | 3.1609 | 4.0778 | 0.6448 |
| não-relacionada | -1.5714 | 4.5595 | 0.7209 |

Tabela 2: Janela Temporal de 300ms-500ms (μ V)

Os 80 valores das médias das janelas de interesse (40 de cada condição) foram testados em uma ANOVA de medidas repetidas, que revelou um valor-p muito próximo do zero. Como predito, então, pela hipótese inicial, há uma diferença entre a condição não-relacionada e a condição relacionada. A deflexão negativa na janela de 300ms a 500ms é claramente visualizada na Figura 4 na linha tracejada, que representa a diferença entre condições.

| | Valor F | Pr > F |
|----------|----------|--------|
| Condição | 108.4796 | 0.0000 |

Tabela 3: ANOVA de Medidas Repetidas na extração de ERPs

Iniciou-se o procedimento baseado na TFD pela decomposição de cada *trial* pela aplicação da Transformada de Fourier ao sinal, usando como *kernel* a Wavelet Complexa de Morlet. Esse procedimento permite que cada *trial* seja decomposto em uma matriz de bandas de frequência i e janelas temporais j . Para cada x_{ij} foi calculada a magnitude M , que é o quadrado no poder calculado para aquele ponto específico. Como discutido na seção 2.2.2, é uma característica do sinal eletrofisiológico oscilar com menos energia em bandas de frequência mais altas, o que acarreta em poder e magnitudes diminutas nessas bandas. Assim, foi realizada a transformação logarítmica em decibéis de cada x_{ij} , tendo como referência

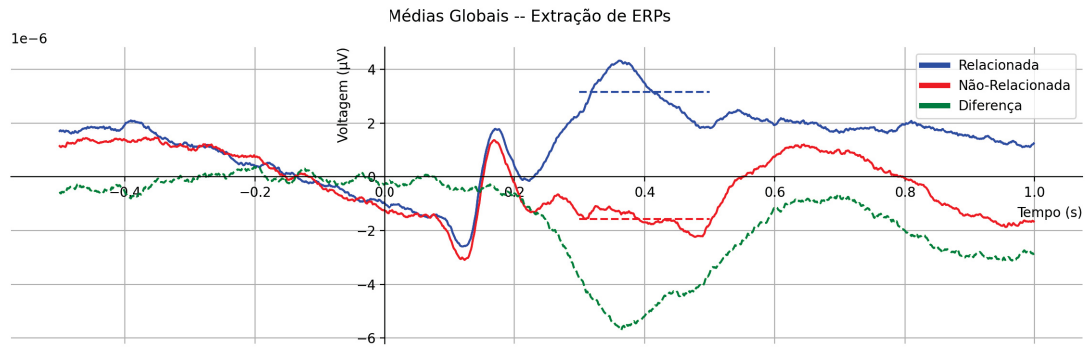


Figura 4: Visualização no domínio temporal das médias globais da Extração de ERPs. As linhas tracejadas representam a média da amplitude de cada condição na janela do N400 (300ms a 500ms). A linha tracejada verde, que é a diferença da condição não-relacionada em relação à relacionada, claramente mostra uma deflexão negativa na janela do N400.

a média de uma janela temporal de 100ms antes do estímulo para cada frequência i . Após o cálculo de uma matriz para cada $trial$, a média de cada x_{ij} por condição foi calculada para sumarização dos dados. Por fim, para visualização foram feitas médias de cada x_{ij} por condição entre todos os 40 participantes do experimento. O resultado é sumarizado na Figura 5. Procedeu-se também o cálculo da diferença entre as condições para melhor visualização dos dados.

Como se pode observar na diferença entre as condições não-relacionada e relacionada, nas bandas delta (2-4Hz) e theta (4-8Hz), o efeito da energia oscilatória do N400, previamente visualizado no domínio temporal, é também visível no domínio do tempo e da frequência, com uma pequena imprecisão temporal. Além disso, é visível uma série de picos positivos de energia na banda gamma (>40Hz). Essa energia parece ao longo de toda janela temporal, mas tem picos próximos aos 500ms e 700ms. Embora não estivesse na hipótese inicial deste artigo, picos de energia de oscilações na banda gamma não são desconhecidos e já foram reportados na literatura sobre o processamento da linguagem em experimentos diferentes do analisado aqui [8, 9, 10]. Procedeu-se, então, três testes de hipótese. Além da hipótese relacionada ao N400, já reportada acima, testou-se a hipótese de que energia na banda gamma é maior quando há um processamento de sentido coerente (condição relacionada) do que quando não há (condição não-relacionada) [8, 9, 10]. Realizaram-se três ANOVAS de Medidas Repetidas em bandas e janelas temporais de interesse: (i) na janela de 400ms, na banda delta, para testar o N400; e (ii) na banda gamma, nas janelas temporais de 450ms e 700ms.

| | Valor F | Pr > F |
|-------------|---------|--------|
| Delta/400ms | 30.2821 | 0.0000 |
| Gamma/450ms | 5.8313 | 0.0205 |
| Gamma/700ms | 7.5063 | 0.0092 |

Tabela 4: ANOVA de Medidas Repetidas na TFD

Os resultados mostram diferenças salientes no três testes de hipótese realizados, descartando portanto H_0 nos três testes.

4.2 Ruído e Sensibilidade

Para testar os dois métodos quanto ao ruído e à sensibilidade do teste de hipótese, em especial no que tange ao Erro Tipo II, dividiu-se a amostra e selecionou-se aleatoriamente dados de 5 participantes, aos quais se foram adicionando um a um mais participantes. Os participantes eram os mesmos nas duas análises e os procedimentos foram os mesmos reportados nas seções anteriores. O objetivo foi estimar quantos participantes são necessários para que o efeito do N400, que é robusto nos dados dos 40 participantes, apareça com um número reduzido de dados. Por uma questão de concisão, reporta-se na Tabela 5 os resultados de 5 em 5 participantes. São reportados nessa tabela os valores do teste F e o valor- p de ANOVAS de medidas repetidas (ANOVAmr) e de regressões lineares usando mínimos quadrados ordinários (RLmq). Os plots dos respectivos grupos encontram-se no Apêndice deste trabalho.

| n | ERP | | TFD | |
|----|--------------|--------------|---------------|---------------|
| | ANOVAmr | RLmqo | ANOVAmr | RLmqo |
| 5 | 4.090(0.113) | 1.144(0.315) | 3.205(0.147) | 4.479(0.0671) |
| 10 | 5.475(0.047) | 1.548(0.231) | 12.870(0.005) | 11.658(0.003) |
| 15 | 18.78(0.007) | 4.13(0.051) | 19.488(0.000) | 12.874(0.001) |
| 20 | 33.99(0.000) | 7.52(0.009) | 12.110(0.002) | 11.016(0.002) |

Tabela 5: Teste de Hipótese com variação no número de participantes (n) aumentando de 5 em 5 (teste F/valor-p).

Embora o valor- p não possa ser considerado como um parâmetro absoluto para rejeição da H_0 , como todos os demais parâmetros são mantidos constantes, vê-se que com menos participantes a análise baseada em TFD identifica o efeito do N400. O ponto crucial é que com 10 participantes, já é possível rejeitar H_0 quando se utiliza a TFD, mas não quando se utiliza a extração de ERPs.

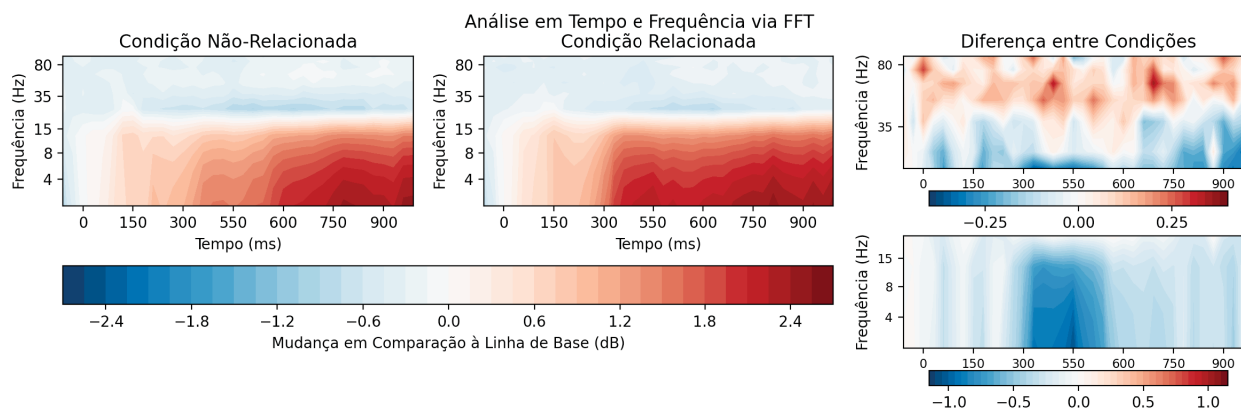


Figura 5: Visualização no Domínio do Tempo e da Frequência por meio de médias globais após Decomposição em Tempo e Frequência (TFD). O primeiro quadro apresenta a condição não-relacionada; o segundo quadro apresenta a condição relacionada; o terceiro quadro superior e inferior apresentam a diferença entre as condições: acima a banda gamma (>40Hz) e abaixo as bandas delta (2-4Hz), beta (4-8Hz) e alpha (8-18Hz) com a magnitude em diferentes escalas.

5 Discussão

A utilização dos dois métodos de análise de dados de EEG feita neste artigo possibilita sua comparação direta e a discussão das suas vantagens e desvantagens. Na seção 4, foram apresentadas as duas análises. A extração de ERPs foi esgotada: a maior parte das recomendações e análises possíveis reportadas na literatura foram utilizadas. Os resultados se coadunam com a hipótese inicial e se visualiza o efeito esperado N400, que é conhecido há 40 anos na literatura sobre processamento da linguagem [4], e há 30 anos no paradigma experimental testado aqui [13]. Na análise baseada em TFD, também foi possível visualizar o efeito do N400. Reproduzir esse efeito era um bom teste para a TFD, uma vez que o N400 é um efeito conhecido em eletrofisiologia do processamento linguístico e esperado nesse paradigma experimental.

Há duas vantagens salientes da TFD sobre a extração de ERPs: (i) a possibilidade de analisar mais frequências e (ii) a necessidade de menos participantes para aumentar a razão sinal/ruído nos dados de EEG. Nos resultados reportados na seção 4, a análise da banda de frequências *gamma*, cujo efeito entre condições apareceu em duas janelas temporais (450ms e 700ms), só é possível com a utilização da TFD. Esse efeito se coaduna com resultados reportados anteriormente na literatura, em que as oscilações em bandas de frequência *gamma* apresentam magnitudes ampliadas em sequências coerentes de palavras em comparação com condições incoerentes [8, 9, 10]. Essa hipótese só pode ser testada com a utilização da TFD, uma vez que, na extração de ERPs, o procedimento de cálculo de médias cancela essas oscilações de frequência alta por serem consideradas ruído a ser eliminado. Além disso, como mostrado na seção 4.2, o teste de hipótese apresenta maior sensibilidade quando é feita a análise baseada em TFD do que quando a extração de ERPs é realizada. Com dados de apenas 10 participantes, tanto a ANOVA de medidas repetidas quanto a regressão linear mostraram o efeito do N400

quando a TFD foi aplicada, mas somente a primeira revelou um valor-*p* abaixo de 0,05 quando a extração de ERPs foi aplicada, ainda assim muito próxima do "valor de corte". Se fosse aplicada qualquer correção para múltiplos testes de hipótese, que são comuns em análises de dados de EEG, como por exemplo a de Bonferroni, os valores provavelmente não estariam abaixo do ponto de corte. Assim, a TFD é mais robusta que a extração de ERPs em termos de aumento da razão/ruído. Ao mesmo tempo, a TFD não elimina a informação de outras frequências, mas antes a decompõe, e assim possibilita análise de outras bandas além das normalmente analisadas na extração de ERPs.

Há, no entanto, uma desvantagem da TFD com relação à extração de ERPs. Há um *trade-off* entre precisão de tempo e de frequência inerente à decomposição baseada na Transformada de Fourier. Essa desvantagem aparece no pico do N400 quando se comparam as Figuras 4 e 5: na extração de ERPs, o pico da onda que representa a diferença entre as condições ocorre um pouco antes dos 400ms; na TFD, o pico de diferença da banda de frequências baixas ocorre aos 550ms. A precisão temporal da extração de ERPs, no entanto, não é verdadeiramente utilizada. Quando se faz o teste de hipótese, por exemplo, se utiliza uma amplitude média de uma janela temporal estabelecida com base na literatura. Não há um consenso sobre como estabelecer essa janela. A precisão temporal, portanto, não é de grande impacto para os estudos baseados na extração de ERPs.

Por fim, há uma crítica levantada pelos teóricos que advogam somente pela extração de ERPs na análise da eletrofisiologia cognitiva: a existência de energia em uma dada frequência pode não ser uma realidade fisiológica, apesar de estar na decomposição gerada pela Transformada de Fourier [2]. Essa ideia surge do fato de que a TFD é uma operação matemática para decompor um conjunto de oscilações em uma combinação de amplitudes, frequências e fases, que foi inicialmente pensada para ondas estacionárias, mas não para ondas eletrofisiológicas

que são altamente transientes. Esse argumento, porém, pode ser refutado de duas formas. Primeiramente, a não-estacionariedade das ondas no mundo físico é um fenômeno extremamente pervasivo em vários domínios. Para citar um exemplo relacionado à linguagem, os sons da voz não são estacionários, uma vez que se o fossem só seria possível ouvir um zunido, e não as vogais e as consoantes que são utilizadas na comunicação. A transiência das ondas é utilizada, portanto, para veicular informação. Segundo, a TFD é uma aproximação matemática de um fenômeno natural, ou seja, um modelo. Como tal, a TFD tem seus pressupostos e é capaz de se aproximar até certo ponto da realidade natural, mas jamais terá uma conexão imediata com ela. Assim, a existência real de oscilações no cérebro e a utilização dessas oscilações para codificar informação cognitiva é sujeita a comprovação ou refutação por meio da reprodutibilidade dos resultados e da utilização de diferentes técnicas para coleta de dados. Até o momento, o que se sabe sobre oscilações neurofisiológicas é que (i) elas ocorrem em todas as espécies; (ii) elas também ocorrem no nível celular em testes *in vitro* e (iii) elas também aparecem em coletas de dados de participantes que têm eletrodos implantados diretamente no cérebro (EEG intracraniana) [3]. Esses três conjuntos de evidências apontam no sentido da realidade psicológica e física das oscilações, ainda que sejam necessários mais dados para ligá-las a fenômenos cognitivos específicos.

6 Conclusão

A importância dos dados de EEG para o estudo neurofisiológico do processamento da linguagem é bem conhecida [6, 9, 1]. Assim, esse artigo se concentrou em fazer uma comparação entre dois grandes grupos de análise de dados de EEG para esse tipo de estudo: a extração de ERPs e a abordagem baseada TFD. Há duas justificativas para se fazer essa comparação: (i) sabendo as vantagens e desvantagens, bem como uma breve compilação dos argumentos a favor e contra cada uma dessas análises, pesquisadores podem fazer uma escolha informada do tipo de abordagem que será utilizada e quais são suas possibilidades e limitações e (ii) o código completo para a realização dessas duas análises pode ser disponibilizado, pois Python é gratuito e *opensource*. A escolha entre uma ou outra análise não será baseada em critérios externos (como o preço da aquisição de *software* que faz ERPs ou TFD), mas sim em critérios científicos. Há ainda muito campo a ser explorado tanto em termos teórico-metodológicos, como por exemplo análises que discutam a localização espacial dos dados de EEG, quanto em termos de implementação, como por exemplo a implementação de certos filtros e interfaces de interação que, pelo que é conhecido, só está disponível em *software* pago. Essas extensões poderão ser exploradas no futuro. No entanto, esse artigo pode ser visto como um passo na direção de tornar diferentes análises de dados de EEG mais acessíveis e assim, com mais acessibilidade, trabalho e criticismo, mais robustas.

Agradecimentos

Gostaria de agradecer ao orientador deste trabalho, professor Paulo Justiniano Ribeiro Junior, por aceitar supervisionar o processo e pelas valiosas discussões. Gostaria também de agradecer à banca que examinou o artigo e aos colegas com os quais compartilhei as ideias apresentadas aqui e o aprendizado do curso.

Referências

- [1] Anna M. Beres. Time is of the Essence: A Review of Electroencephalography (EEG) and Event-Related Brain Potentials (ERPs) in Language Research. *Applied Psychophysiology Biofeedback*, 42, 2017.
- [2] Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2014.
- [3] Mike X. Cohen. *Analyzing neural time series data: Theory and practice*. MIT Press, 2014.
- [4] Marta Kutas and Stephen Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427), 1980.
- [5] Marta Kutas and Cyma K. Van Petten. Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher, editor, *Handbook of Psycholinguistics*, pages 83–143. Academic Press, San Diego, CA, US, 1994.
- [6] Marta Kutas, Cyma K. Van Petten, and Robert Kluender. Psycholinguistics electrified II: 1994–2005. In M. Traxler and M. A. Gernsbacher, editors, *Handbook of Psycholinguistics*, page 659–724. Elsevier, New York, NJ, US, 2006.
- [7] Joseph Dien. Best practices for repeated measures ANOVAs of ERP data: Reference, regional channels, and robust ANOVAs. *International Journal of Psychophysiology*, 111, 2017.
- [8] Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669):438–441, 2004.
- [9] Marcel Bastiaansen and Peter Hagoort. Frequency-based segregation of syntactic and semantic unification during online sentence level language comprehension. *Journal of cognitive neuroscience*, 27(11):2095–2107, 2015.
- [10] Yifei He, Arne Nagels, Matthias Schlesewsky, and Benjamin Straube. The role of gamma oscillations during integration of metaphoric gestures and abstract speech. *Frontiers in Psychology*, 9:1348, 2018.
- [11] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks,

- Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7:267, 2013.
- [12] Emily S. Kappenman, Jaclyn Farrens, Wendy Zhang, Andrew X. Stewart, and Steven J. Luck. ERP CORE: An Open Resource for Human Event-Related Potential Research [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/4azqm>, 2020.
- [13] Phillip J. Holcomb and Helen J. Neville. Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes*, 5(4):281–312, 1990.
- [14] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.
- [15] Guido Van Rossum and Fred L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- [16] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

A Apêndice

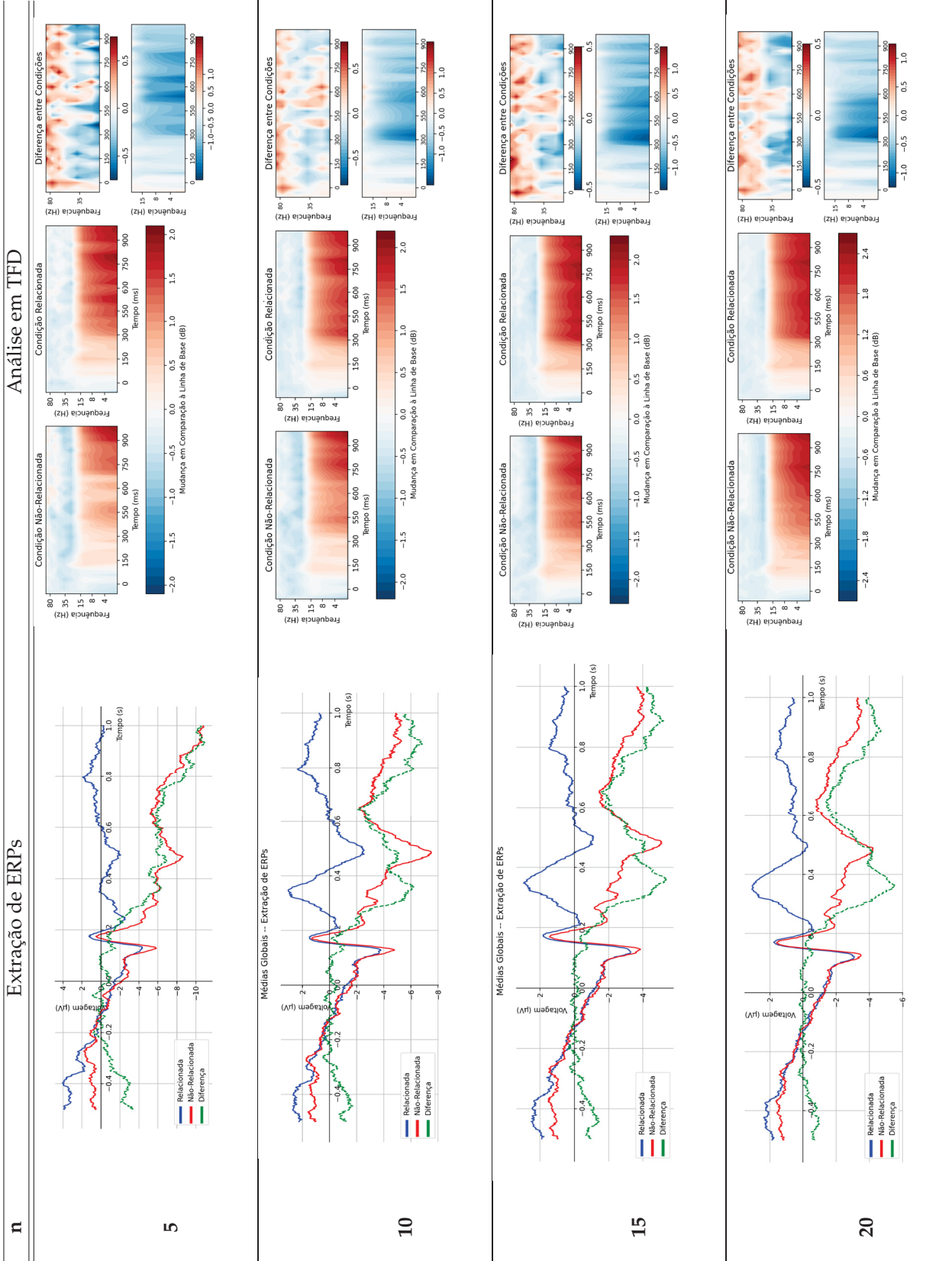


Tabela 6: Comparação da visualização da Extração de ERPs e da análise em TFD com diferentes números de participantes.